ELSEVIER

CrossMark

Journal of Clinical Epidemiology 70 (2016) 106–110

Journal of Clinical Epidemiology

An algorithm was developed to assign GRADE levels of evidence to comparisons within systematic reviews

Alex Pollock[a,*], Sybil E. Farmer[a], Marian C. Brady[a], Peter Langhorne[b], Gillian E. Mead[c], Jan Mehrholz[d], Frederike van Wijck[e], Philip J. Wiffen[f]

[a]*Nursing, Midwifery and Allied Health Professions Research Unit, Glasgow Caledonian University, Buchanan House, Cowcaddens Road, Glasgow G4 0BA, UK*

[b]*Academic Section of Geriatric Medicine, Level 2, New Lister Building, Glasgow Royal Infirmary, Glasgow G31 2ER, UK*

[c]*Centre for Clinical Brain Sciences, Royal Infirmary of Edinburgh, Little France Crescent, Edinburgh EH16 5NN, UK*

[d]*Wissenschaftliches Institut, Private Europäische Medizinische Akademie der Klinik Bavaria in Kreischa GmbH, An der Wolfsschlucht 1–2, 01731 Kreischa, Germany*

[e]*Institute for Applied Health Research and the School of Health and Life Sciences, Glasgow Caledonian University, Cowcaddens Road, Glasgow G4 0BA, UK*

[f]*Pain Research and Nuffield Department of Clinical Neurosciences (Nuffield Division of Anaesthetics), University of Oxford, Churchill Hospital, Oxford OX3 7LE, UK*

Accepted 10 August 2015; Published online 1 September 2015

Abstract

Objectives: One recommended use of the Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach is supporting quality assessment of evidence of comparisons included within a Cochrane overview of reviews. Within our overview, reviewers found that current GRADE guidance was insufficient to make reliable and consistent judgments. To support our ratings, we developed an algorithm to grade quality of evidence using concrete rules.

Methods: Using a pragmatic, exploratory approach, we explored the challenges of applying GRADE levels of evidence and developed an algorithm to applying GRADE levels of evidence in a consistent and transparent approach. Our methods involved application of algorithms and formulas to samples of reviews, expert panel discussion, and iterative refinement and revision.

Results: The developed algorithm incorporated four key criteria: number of participants, risk of bias of trials, heterogeneity, and methodological quality of the review. A formula for applying GRADE level of evidence from the number of downgrades assigned by the algorithm was agreed.

Conclusion: Our algorithm which assigns GRADE levels of evidence using a set of concrete rules was successfully applied within our Cochrane overview. We propose that this methodological approach has implications for assessment of quality of evidence within future evidence syntheses.

Keywords: GRADE; Review; Overview; Algorithm; Methodology; Quality of evidence

1. Introduction

Cochrane overviews of reviews (Cochrane overviews) compile evidence from multiple systematic reviews of interventions into one accessible document [1]. An essential part of an overview is the assessment of the quality of evidence arising from the included reviews, and the Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach [2,3] is the framework recommended by the Cochrane Handbook to facilitate transparent rating of quality of evidence [1]. Although principally developed for grading recommendations during guideline development, the GRADE system for assessing the quality of a body of evidence has gained widespread use within systematic reviews. Journal of Clinical Epidemiology has previously invited researchers to share their experiences of using GRADE [4].

We have recently published a Cochrane overview of reviews relating to interventions to improve upper limb (arm) function after stroke [5]. Within this we identified, from 40 included systematic reviews, 127 comparisons with

Funding: The Cochrane overview was supported by a project grant (CZH/4/854) from the Chief Scientist Office (CSO), part of the Scottish Government Health and Social Care Directorate.

Conflict of interest: None.

* Corresponding author. Tel.: +44-(0)141-331-8100; fax: +44-(0)141-331-8101.

E-mail address: alex.pollock@gcu.ac.uk (A. Pollock).

http://dx.doi.org/10.1016/j.jclinepi.2015.08.013

What is new?

Key findings

- We developed an algorithm which provided a set of concrete rules for grading the quality of evidence within our Cochrane overview.
- The algorithm enables efficient, transparent, reliable grading of quality of evidence.

What this adds to what was known?

- Our experience adds to the debate in the literature about the advantages and disadvantages of using concrete rules to apply Grading of Recommendations Assessment, Development and Evaluation levels of evidence.

What is the implication and what should change now?

- We have demonstrated that it is possible to develop an algorithm based on detailed subjective judgment of what is most important to a specific body of evidence.
- This methodological approach has implications for assessment of quality of evidence within future evidence syntheses, which should build on our work in this area.

relevant outcomes [5]. We planned to grade the evidence for each relevant comparison using the GRADE definitions of high, moderate, low, or very low quality of evidence [2,6–8], using definitions of these criteria from published literature on GRADE [9]. However, Berkman et al. [10] have previously found poor agreement on grading strength of evidence within systematic reviews using GRADE, even among experienced systematic reviewers. Berkman et al. [10] concluded that more specific guidance was required to assist reviewers in judging grade of evidence. In contrast to Balshem et al. [6], who warned that a ''mechanistic approach'' would obviate the essential component of subjective judgment, Berkman et al. [10] proposed exploring the ''advantages and disadvantages of concrete rules for reaching strength of evidence grades.''

In this brief report, we share our experience of developing and applying concrete rules to grading quality of evidence of the 127 comparisons included in our overview and discuss whether this was an advantageous approach.

2. Methods

The methods we used were exploratory, developed iteratively and pragmatically, to first explore the challenges of applying GRADE levels of evidence, and second, develop a systematic, consistent, and transparent approach to applying GRADE to efficiently grade the quality of evidence within these 127 comparisons.

2.1. Exploration of applying GRADE levels of evidence

We explored the process of applying GRADE levels of evidence to a sample of four purposively selected reviews, chosen to reflect a range of different types, format, and quality of reviews. For the comparisons within these four reviews, two independent, experienced, reviewers with clinical knowledge relating to the topic assessed risk of bias, publication bias, indirectness, imprecision, and inconsistency and applied the GRADE levels of evidence based on their considered judgment. The reviewers first familiarized themselves with published materials relating to best practices in using the GRADE framework, in particular the 2011 Journal of Clinical Epidemiology 20-part series [3]. The applied grades of evidence of the two reviewers were compared and discussed. No statistical assessment of agreement between reviewers was carried out, as the aim of this exercise was simply to explore the application of GRADE to inform subsequent decisions regarding the methods we would adopt for the overview.

2.2. Development of the initial algorithm

We used an expert panel approach; the overview author team and Professor Wiffen, who was invited to share his methodological expertise and experience relating to Cochrane overviews, attended an all-day meeting. We had substantive discussions over the application of GRADE criteria, and the reviewers' perceived challenges associated with the process. We agreed to develop a set of objective criteria for determining whether the evidence relating to each comparison was of high, moderate, low, or very low quality. The team discussed in depth which objective criteria were perceived to be most relevant to a judgment of quality of the particular evidence synthesized within this overview, and specific objective criteria perceived to support a judgment of ''high,'' ''moderate,'' or ''low'' were explored. An initial algorithm for assigning downgrades to the level of evidence was drafted based on these criteria.

2.3. Testing and refining of algorithm

The initial developed algorithm was applied to a convenience sample (alphabetical selection) of five reviews (containing 43 comparisons) identified for inclusion in the overview. The assigned level of evidence based on this initial algorithm was explored and compared to previous, independently applied, subjective judgments of all overview authors. An iterative process, involving discussion between all overview authors, was then carried out, exploring the impact of the ''weighting'' assigned to each of the criteria included in the algorithm. Discussion and further

Table 1. Key criteria considered to be of most importance to our overview and used to make objective decisions relating to the GRADE level of evidence within the developed algorithm

Criteria considered important	Rationale	Further information
1. Number of participants	Evidence demonstrates an association between trial sample size and treatment effects.	Dechartres et al. [11] found that trials with 1,000 patients or more had, on average, a treatment effect which was 48% smaller than trials with less than 50 patients. Furthermore, Dechartres et al. [12] found that the treatment outcome estimates from pooled meta-analyses within systematic reviews were generally larger than the results of a meta-analysis of the largest trials only, particularly for subjective outcomes.
2. Risk of bias of trials	Assessment of risk of bias of trials is central to ensuring quality of systematic reviews [13,14], and trials at high risk of bias are likely to report exaggerated effect size [15].	For the body of evidence which we were assessing, we considered that method of randomization and blinding of outcome assessor were of the highest importance in assessing risk of bias.
3. Heterogeneity	Statistical measures of heterogeneity within meta-analyses can reflect inconsistency in results, and I^2 provides a recognized measure of statistical heterogeneity.	With an I^2 greater than 75% potentially being ''considerable'' [16], we decided that this provided a simple, transparent, and objective criterion for assessment of inconsistency.
4. Methodological quality of the review	Methodological quality of the review from which data are extracted is a key factor in the quality of evidence.	We had previously assessed quality of each included review using the AMSTAR [17,18]. Reviewers agreed that questions relating to (1) a priori research design, (2) search characteristics, (3) independence of study selection, and (4) data extraction were of the greatest importance to the methodological quality of the reviews within this overview, rather than using the total AMSTAR ''score'' (see Appendix at www.jclinepi.com for details of these questions).

Abbreviation: GRADE, Grading of Recommendations Assessment, Development and Evaluation; AMSTAR, the AMSTAR quality assessment tool [17,18].

refinement continued until all authors agreed on a revised algorithm, judged to assign an appropriate number of downgrades in response to different objective data.

2.4. Development of formula to assign GRADE level of evidence

We created and applied four different versions of a formula to assign levels of evidence based on numbers of downgrades to each of the 43 sample comparisons, and the resulting levels of evidence were explored and discussed by the overview authors. Discussion continued until all authors reached consensus on the number of downgrades which should equate to the GRADE levels of evidence of high, moderate, low, and very low.

3. Results

During the exploration of applying GRADE to a sample of reviews, both reviewers communicated that they felt there was a great deal of subjectivity involved in the assessment of GRADE levels of evidence and that the GRADE criteria were poorly defined and difficult to apply. The reviewers both recounted that they lacked sufficient guidance to make judgments confidently, and agreement between reviewers on GRADE levels of evidence was inconsistent. Furthermore, both reviewers independently commented that their decisions to downgrade the level of evidence were largely being influenced by objective data that had been extracted for each comparison. In particular, both reviewers reported using specific data to inform their decision. These data commonly included the results of the assessment of risk of bias (on the Cochrane risk of bias tool), measures of heterogeneity (e.g., I^2), and the volume of evidence (number of trials and included participants).

To develop the initial algorithm, it was agreed that the key criteria which were of most importance to our overview and which should be used to make objective decisions relating to the GRADE level of evidence were number of participants; risk of bias of trials, specifically in relation to randomization and blinded outcome assessment; heterogeneity; and methodological quality of the review. The rationale and supporting evidence for agreement over these four criteria are provided in Table 1. Table 2 illustrates the final algorithm which was agreed to be most appropriate for determining the number of downgrades to levels of evidence in reviews. This algorithm results in a maximum of six downgrades for each assessed comparison; Table 3 illustrates the agreed formula for applying GRADE level of evidence from the number of assigned downgrades.

4. Discussion

We developed an algorithm which provided a set of concrete rules for grading quality of evidence. This algorithm

Table 2. Algorithm for determining "downgrades" to levels of evidence in reviews

Area assessed	Imprecision	Risk of bias (trial quality)	Inconsistency	Risk of bias (review quality)
Method of assessment	Number of participants within pooled analysis	Proportion of participants included in the pooled analysis judged to have low ROB for randomization and observer blinding	Heterogeneity, assessed by I^2 statistic	Responses to AMSTAR questions 1–4 (covering a priori research design, search characteristics, independence of study selection, and data extraction)
No downgrade (no serious limitations)	≥200	≥75% of participants have low ROB	$I^2 \leq 75\%$	4/4 are all "yes" (i.e., low ROB)
Downgrade 1 level (serious limitations)	100–199	<75% of participants have low ROB	$I^2 > 75\%$	3/4 are "yes" and 1 is "unclear" or "no" on AMSTAR
Downgrade 2 levels (very serious limitations)	1–99			< 3/4 are "yes" and remainder are "unclear" or "no" on AMSTAR
Notes		If ROB for individual trials was not reported within the review, we were conservative and assumed that less than 75% of participants had low ROB.	If only one trial contributed to analysis, no downgrade; if I^2 not reported, assumed to be greater than 75%.	

Abbreviations: ROB, risk of bias; AMSTAR, the AMSTAR quality assessment tool [17,18].

was developed specifically for our Cochrane overview [5] and with careful consideration of the type of evidence included with this overview. With limited published guidance to aid decisions, the objective criteria, and the "cutoffs" within these criteria, were agreed by the overview team, using an iterative process of exploration and discussion. Undoubtedly, the selection of different methodological criteria, and cutoffs within these criteria, will have impacted on the GRADE levels allocated to evidence within this overview. However, the "weightings" that our methods gave to different methodological criteria were considered in detail by the overview team, and care was taken to ensure that the resultant objectively determined GRADE levels reflected overview authors' more subjective views of the quality of the evidence.

Application of this algorithm to all 127 comparisons within our overview enabled us to efficiently compute the GRADE level of evidence based on objective data extracted from the included reviews. The algorithm afforded a transparent and reproducible system, which avoided the need for subjective decision making, and we believe this is a key strength of our overview. However, we are not advocating that our specific algorithm is directly applied to the assessment of quality of evidence within other overviews or in the assessment of other bodies of evidence. Rather we argue that this approach may have direct relevance and implications to authors of other overviews or those involved in assessing other bodies of evidence. Our algorithm was developed specifically to reflect what we judged to be important to the quality of this particular body of evidence. Both the criteria we included and the cutoffs we applied were specifically—one might even argue "subjectively"—selected to reflect our expert opinions and considerations relating to what we perceived to be high-, moderate-, or low-quality evidence relating to interventions to improve upper limb function after stroke. Thus, although we have been mechanistic in our application of our algorithm, we were not mechanistic in its development. The algorithm was developed based on detailed subjective judgment, informed by in-depth understanding of the topic area, of what was most important to the quality of this specific body of evidence.

Table 3. Formula for applying GRADE level of evidence from number of downgrades determined using the algorithm

GRADE level of evidence	Number of downgrades (derived from objective assessment)
High	0 downgrade
Moderate	1 or 2 downgrades
Low	3 or 4 downgrades
Very low	5 or 6 downgrades

Abbreviation: GRADE, Grading of Recommendations Assessment, Development and Evaluation.

5. Conclusion

We developed an algorithm, which assigned GRADE levels of evidence using a set of concrete rules, and successfully applied this within our Cochrane overview. Our algorithm was developed to assess the specific body of evidence synthesized within our overview; however, we believe that our approach to the application of GRADE levels of evidence may have widespread applicability. We propose that this methodological approach has implications for assessment of quality of evidence within future evidence syntheses, which will be able to build on our work in this area.

Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.jclinepi.2015.08.013.

References

[1] Becker LA, Oxman AD. Chapter 22: overviews of reviews. In: Higgins JPT, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]. The Cochrane Collaboration; 2011. Available at: www.cochrane-handbook.org Accessed January 6, 2015.

[2] Guyatt GH, Oxman AD, Vist G, Kunz R, Falck-Ytter Y, for the GRADE Working Group. Rating quality of evidence and strength of recommendations GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008;336:924−6.

[3] Guyatt GH, Oxman AD, Schünemann HJ, Tugwell P, Knotterus A. GRADE guidelines: a new series of articles in the Journal of Clinical Epidemiology. J Clin Epidemiol 2011a;64:380−2.

[4] Straus SE, Shepperd S. Challenges in guideline methodology. J Clin Epidemiol 2011;64:347−8.

[5] Pollock A, Farmer SE, Brady MC, Langhorne P, Mead GE, Mehrholz J, et al. Interventions for improving upper limb function after stroke. Cochrane Database Syst Rev 2014;CD010820.

[6] Balshem H, Helfand M, Schunemann HJ, Oxman AD, Kunz R, Brozek J, et al. GRADE guidelines 3: rating the quality of evidence—introduction. J Clin Epidemiol 2011;64:401−6.

[7] Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, et al. GRADE guidelines: 1. Introduction—GRADE evidence profiles and summary of findings tables. J Clin Epidemiol 2011b;64:383−94.

[8] Guyatt GH, Oxman AD, Vist G, Kunz R, Brozek J, Alonso-Coello P, et al. GRADE guidelines 4: rating the quality of evidence—risk of bias. J Clin Epidemiol 2011c;64:407−15.

[9] GRADE Working Group. Available at: http://www.gradeworkinggroup.org Accessed January 6, 2015.

[10] Berkman ND, Lohr KN, Morgan LC, Kuo T-M, Morton SC. Interrater reliability of grading strength of evidence varies with the complexity of the evidence in systematic reviews. J Clin Epidemiol 2013;66:1105−1117.e1.

[11] Dechartres A, Trinquart L, Boutron I, Ravaoud P. Influence of trial sample size on treatment effect estimates: meta-epidemiological sample. BMJ 2013;346:f2304.

[12] Dechartres A, Altman D, Trinquart L, Boutron I, Ravaoud P. Association between analytic strategy and estimates of treatment outcomes in meta-analyses. JAMA 2014;312:623−30.

[13] Higgins JPT, Altman DG, Sterne JAC. Chapter 8: assessing risk of bias in included studies. In: Higgins JPT, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 (updated March 2011). The Cochrane Collaboration; 2011. Available at www.cochrane-handbook.org Accessed January 29, 2015.

[14] Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA Statement. PLoS Med 2009;6:e1000097.

[15] Savovic J, Jones HE, Altman DG, Harris RJ, Juni P, Pildal J, et al. Influence of reported study design characteristics on intervention effect estimates from randomized, controlled trials. Ann Intern Med 2012;157:429−38.

[16] Schünemann H, Brożek J, Guyatt G, Oxman A, editors. GRADE handbook for grading quality of evidence and strength of recommendations. The GRADE Working Group; 2013. Available at: www.guidelinedevelopment.org/handbook Accessed January 28, 2015.

[17] Shea BJ, Grimshaw JM, Wells GA, Boers M, Andersson N, Hamel C, et al. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. BMC Med Res Methodol 2007;7:10.

[18] Shea BJ, Hamel C, Wells GA, Bouter LM, Kristjansson E, Grimshaw J, et al. AMSTAR is a reliable and valid measurement tool to assess the methodological quality of systematic reviews. J Clin Epidemiol 2009;62:1013−20.